Available online at www.sciencedirect.com

SCIENCE DIRECT®

Advances in Applied Mathematics 34 (2005) 65–70

ADVANCES IN Applied Mathematics

www.elsevier.com/locate/yaama

De Bruijn covering codes with arbitrary alphabets

V. Vu [1]

Department of Mathematics, UCSD, La Jolla, CA 92093-0112, USA

Received 30 September 2003; accepted 4 May 2004

Available online 6 October 2004

Abstract

Let Ω be a set of q symbols and $\Omega^n = \{x_1 \dots x_n \mid x_i \in \Omega\}$. We prove that for any fixed q and R, there is a de Bruijn covering code of radius R of length $O(\frac{q^n}{\binom{n}{R}} \ln n)$, answering a question of Chung and Cooper.

1. Introduction

Let Ω be a set of q symbols and $\Omega^n = \{x_1 \dots x_n \mid x_i \in \Omega\}$. An element of Ω^n is a *codeword* of length n with respect to the alphabet Ω. The most popular case is when $\Omega = \{0, 1\}$ and in this case we talk about binary codes. In this paper, we can assume, without loss of generality, that $\Omega = \{0, 1, \dots, q-1\}$.

Let $x = x_1 \dots x_n$ and $y = y_1 \dots y_n$ be two codewords. The Hamming distance between x and y is the number of coordinate i where $x_i \neq y_i$. A subset $X \subset \Omega^n$ is a *covering code* of radius R if for every codeword $y \in \Omega^n$, there is a codeword $x \in X$ such that the Hamming distance between x and y is at most R. Here and later, we assume that q and R are fixed and n is sufficiently large. The asymptotic notation will be used under the assumption that

E-mail address: vanvu@ucsd.edu.

URL: http://www.math.ucsd.edu/vanvu/.

[1] The author is supported by an A. Sloan fellowship, an NSF Career Award and NSF grant DMS-0200357.

0196-8858/$ – see front matter
doi:10.1016/j.aam.2004.05.004

$n \to \infty$. It is easy to see that for any fixed x, the number of codewords of distance at most R from x is

$$V(n, R) = \sum_{i=0}^{R} \binom{n}{i} (q-1)^i = \big(1 + o(1)\big)\binom{n}{R}(q-1)^R.$$

It follows that any covering code of radius R should have at least

$$\frac{|\boldsymbol{\Omega}^n|}{V(n, R)} = \frac{q^n}{V(n, R)} \tag{1}$$

elements. It was shown recently that this lower bound is sharp, up to a factor roughly $eR \ln R$ (see [5] for the precise statement). For more information about covering codes, we refer to [2].

In a recent paper [1], Cooper and Chung introduced the notion of *de Bruijn covering code*. A sequence $X = x_1 \dots x_m$, $x_i \in \boldsymbol{\Omega}$, is a de Bruijn covering code of radius R if the m codewords $x_1 \dots x_n, x_2 \dots x_{n+1}, \dots, x_m \dots x_{m+n-1}$ form a covering code of radius R, where $x_{m+i} = x_i$. This definition is motivated by the notion of de Bruijn cycles and we say that m is the *length* of the code. Combining tools from linear algebra, field theory and probability, they showed that for special values of q, one can always find a relatively short de Bruijn covering code [1].

Theorem 1.1. *For any fixed q which is a prime power and any fixed R, there is a de Bruijn covering code of radius R with length $m = O(\frac{q^n}{V(n,R)} \ln n)$.*

Chung and Cooper asked whether Theorem 1.1 can be extended for arbitrary q (see the last section of [1]). The goal of this note is to affirmatively answer this question. We prove

Theorem 1.2. *For any fixed q and R, there is a de Bruijn covering code of radius R with length $m = O(\frac{q^n}{V(n,R)} \ln n)$.*

Our proof uses combinatorial arguments and Janson–Suen inequality. Linear algebra and field theory are not required and so we do not need the prime power assumption for q.

2. Proof of Theorem 1.2

For a sequence $X = x_1 \dots x_m$, define

$$S_X = \{s_1 = x_1 \dots x_n,\ s_2 = x_2 \dots x_{n+1},\ s_m = x_m \dots x_{m+n-1}\}.$$

Let C_X denote the set of codewords which are of distance at most R from S_X. Assume that C_X contains all but l codewords in $\boldsymbol{\Omega}^n$. If $v_1, \dots, v_l$ are the codewords not in C_X, then the sequence

$$X' = X x_1 \dots x_{n-1} v_1 \dots v_l,$$

obtained by first extending X by $x_1, \ldots, x_{n-1}$ and next concatenating the extended sequence with $v_1, \ldots, v_l$ is clearly a de Bruijn covering code of radius R. The length of X' is $m + (n-1) + ln$. Thus, in order to prove Theorem 1.2, it suffices to show

Claim 2.1. *There is a sequence X of length $O(\frac{q^n}{V(n,R)} \ln n)$ such that C_X contains all but at most $l = \frac{q^n}{V(n,R)n}$ codewords in $\boldsymbol{\Omega}^n$.*

A random sequence $x_1 x_2 \ldots x_m$ is constructed as follows. For each position i, x_i takes a value from the alphabet $\boldsymbol{\Omega} = \{0, 1, \ldots, q-1\}$ with equal probability $1/q$. We are going to show that a random sequence X of length $m = c\frac{q^n}{V(n,R)} \ln n$, where c is an appropriate constant, satisfies the statement of Claim 2.1 with positive probability.

For a codeword v, let A_v be the event that v is not contained in C_X. By linearity of expectation, the expectation of the number of codewords not contained in C_X is $\sum_{v \in \boldsymbol{\Omega}^n} \mathbf{P}(A_v)$. By symmetry, this number is equal to $q^n \mathbf{P}(A_{\underline{0}})$, where $\underline{0} = 0 \ldots 0$. Therefore, in order to prove Claim 2.1, we need only show that

$$\mathbf{P}(A_{\underline{0}}) \leqslant \frac{1}{V(n,R)n}. \tag{2}$$

Let N the set of codewords of weight at most R (the weight of a codeword v is the number of non-zero symbols in v); these are the codewords with distance at most R from $\underline{0}$. The event $A_{\underline{0}}$ occurs if and only if S_X does not intersect N, namely, s_i does not belong to N for all $i = 1, \ldots, m$. Denote by B_i the event that s_i belongs to N, we have

$$\mathbf{P}(A_{\underline{0}}) = \mathbf{P}\left(\bigwedge_{i=1}^{m} \overline{B_i}\right). \tag{3}$$

We write $i \sim j$ if the intervals s_i and s_j overlap. This relation $\sim$ defines a graph on I with the following property. Let J_1 and J_2 be two disjoint subsets of I such that there are no $i_1 \in J_1$ and $i_2 \in J_2$ where $i_1 \sim i_2$. Let A^1 be any Boolean function of the events $B_i, i \in J_1$, and let A^2 be any Boolean function of the events $B_i, i \in J_2$. Then A^1 and A^2 are independent.

Let $\mu = \sum_{i=1}^{m} \mathbf{P}(B_i)$ and $\Delta = \sum_{i \sim j} \mathbf{P}(B_i \wedge B_j)$ and $\delta = \max_i \sum_{j \sim i} \mathbf{P}(B_j)$. We are going to use the following inequality, due to Janson (Theorem 3 in [3]), which is an variant of an earlier result of Suen [4]

Lemma 2.2. *Under the above notation*

$$\mathbf{P}\left(\bigwedge_{i=1}^{m} \overline{B_i}\right) \leqslant \exp\left(-\min\left(\frac{\mu^2}{8\Delta}, \frac{\mu}{2}, \frac{\mu}{6\delta}\right)\right).$$

To conclude the proof, it remains to estimate μ, Δ and δ. For μ, it is clear that

$$\mu = m\mathbf{P}(B_1) = m\frac{|N|}{q^n} = m\frac{V(n,R)}{q^n}. \tag{4}$$

Next, for each i there are exactly $2(n-1)$ j such that $i \sim j$, so

$$\delta = 2(n-1)\mathbf{P}(B_1) < 2n\frac{V(n,R)}{q^n} = o(1).$$

We will show that

$$\Delta \leqslant \big(1+o(1)\big)m\frac{\binom{n}{R}(q-1)^{R-1}}{q^n} = \frac{(1+o(1))}{q-1}m\frac{V(n,R)}{q^n} = \frac{(1+o(1))}{q-1}\mu. \tag{5}$$

Assuming (5), we conclude the proof as follows. Since $\delta = o(1)$,

$$\min\left(\frac{\mu^2}{8\Delta}, \frac{\mu}{2}, \frac{\mu}{6\delta}\right) = \min\left(\frac{\mu^2}{8\Delta}, \frac{\mu}{2}\right) \geqslant c(q)\mu,$$

where $c(q) = 1/2$ for $q \geqslant 5$ and $c(q) = (q-1)/8$ for $q < 5$. Lemma 2.2 yields

$$\mathbf{P}\left(\bigwedge_{i=1}^{m}\overline{B_i}\right) \leqslant \exp\big(-\big(1+o(1)\big)c(q)\mu\big). \tag{6}$$

Given (4), one can find a number

$$m = \frac{1+o(1)}{c(q)}\frac{q^n}{V(n,R)}\ln\big(nV(n,R)\big) = O\left(\frac{q^n}{V(n,R)}\ln n\right)$$

so that the exponent of the right-hand side of (6) is at most $-\ln(nV(n,R))$. It follows that the right-hand side of (6) is upper bounded by $\exp(-\ln(nV(n,R))) = \frac{1}{nV(n,R)}$, proving Claim 2.1.

It remains to verify (5). To do this, notice that

$$\Delta = \sum_{i=1}^{m}\sum_{j\geqslant i, j\sim i}\mathbf{P}(B_i \wedge B_j) = m\sum_{j\geqslant 1, j\sim 1}\mathbf{P}(B_1 \wedge B_j) = m\sum_{j=2}^{n}\mathbf{P}(B_1 \wedge B_j).$$

For $2 \leqslant j \leqslant n$, the two intervals s_1 and s_j share a common subinterval s_{1j} of length $n+j-1$. Let $s_1' = s_1 \backslash s_{1j}$ and $s_j' = s_j \backslash s_{1j}$ and k_1, k_2, k_3 be the number of non-zero elements in s_{1j}, s_1' and s_j', respectively. The event $B_1 \wedge B_j$ holds if and only if both k_1+k_2 and k_1+k_3 are at most R. Denoting by K the set of triples (k_1,k_2,k_3) satisfying this property, we have

$$\mathbf{P}(B_1 \wedge B_j) = \sum_{(k_1,k_2,k_3)\in K}\frac{\binom{n-j+1}{k_1}\binom{j-1}{k_2}\binom{j-1}{k_3}(q-1)^{k_1+k_2+k_3}}{q^{n+j-1}}.$$

Therefore,

$$\begin{aligned}\sum_{j=2}^{n}\mathbf{P}(B_1\wedge B_j) &= \sum_{j=2}^{n}\sum_{(k_1,k_2,k_3)\in K}\frac{\binom{n-j+1}{k_1}\binom{j-1}{k_2}\binom{j-1}{k_3}(q-1)^{k_1+k_2+k_3}}{q^{n+j-1}}\\ &= \sum_{(k_1,k_2,k_3)\in K}\sum_{j=2}^{n}\frac{\binom{n-j+1}{k_1}\binom{j-1}{k_2}\binom{j-1}{k_3}(q-1)^{k_1+k_2+k_3}}{q^{n+j-1}}\\ &\leqslant \sum_{(k_1,k_2,k_3)\in K}\sum_{j=2}^{\infty}\frac{\binom{n}{k_1}\binom{j-1}{k_2}\binom{j-1}{k_3}(q-1)^{k_1+k_2+k_3}}{q^{n+j-1}}\\ &= \sum_{(k_1,k_2,k_3)\in K}S(k_1,k_2,k_3).\end{aligned}$$

If $k_1 = R$, then $k_2 = k_3 = 0$ (since $k_1 + k_2$ and $k_1 + k_3$ are both at most R). In this case

$$S(R,0,0) = \sum_{j=2}^{\infty}\frac{\binom{n}{R}(q-1)^R}{q^{n+j-1}} = \frac{\binom{n}{R}(q-1)^R}{q^n}\sum_{j=2}^{\infty}\frac{1}{q^{j-1}} = \frac{\binom{n}{R}(q-1)^{R-1}}{q^n}.$$

We are going to show that if $k_1 < R$, then $S(k_1,k_2,k_3) = o\big(\binom{n}{R}/q^n\big)$, regardless the values of k_2 and k_3. Observe that

$$\begin{aligned}S(k_1,k_2,k_3) &= \sum_{j=2}^{\infty}\frac{\binom{n}{k_1}\binom{j-1}{k_2}\binom{j-1}{k_3}(q-1)^{k_1+k_2+k_3}}{q^{n+j-1}}\\ &= \frac{\binom{n}{k_1}}{q^n}\sum_{j=2}^{\infty}\frac{\binom{j-1}{k_2}\binom{j-1}{k_3}(q-1)^{k_1+k_2+k_3}}{q^{j-1}}\\ &\leqslant \frac{\binom{n}{k_1}}{q^n}\sum_{j=2}^{\infty}\frac{j^{k_2+k_3}(q-1)^{k_1+k_2+k_3}}{q^{j-1}}\\ &= O\left(\frac{\binom{n}{k_1}}{q^n}\right),\end{aligned}$$

since the series $\sum_{j=2}^{\infty} j^{k_2+k_3}(q-1)^{k_1+k_2+k_3}/q^{j-1}$ converges. On the other hand, if $k_1 < R$, then $\binom{n}{k_1}/q^n = o\big(\binom{n}{R}/q^n\big)$, proving the claim.

We can now conclude that $\sum_{j\sim 1}\mathbf{P}(B_1\wedge B_j) = (1+o(1))\binom{n}{R}(q-1)^{R-1}/q^n$. Thus

$$\Delta \leqslant \big(1+o(1)\big)m\frac{\binom{n}{R}(q-1)^{R-1}}{q^n} = \frac{(1+o(1))}{q-1}m\frac{V(n,R)}{q^n} = \frac{(1+o(1))}{q-1}\mu,$$

as claimed in (5). Our proof is complete.

Remarks. One can have a better value for $c(q)$ in the case $q \geqslant 3$ by applying Suen inequality [4] rather than Lemma 2.2. Suen's inequality would enable one to set $c(q) = (q-2)/(q-1)$. For q being a prime power, Chung and Cooper showed that one can set $c(q) = 1$. We believe that with extra works, one would be able to set $c(q) = c/q$, where c is an absolute constant not depending on q. The real problem, however, is to determine whether the function $\ln n$ in Theorems 1.1 and 1.2 can be replaced by a smaller function.

Acknowledgment

We thank F. Chung for communicating the problem and pointing out reference [1].

References

[1] F. Chung, J. Cooper, De Bruijn cycles for covering codes, submitted for publication.
[2] G. Cohen, I. Honkala, S. Litsyn, A. Lobstein, Covering Codes, North-Holland Math. Library, vol. 54, North-Holland Publishing Co., Amsterdam, 1997.
[3] S. Janson, New versions of Suen's correlation inequality, in: Proceedings of the Eighth International Conference "Random Structures and Algorithms", Poznan, 1997, Random Structures Algorithms 13 (3–4) (1998) 467–483.
[4] S. Suen, A correlation inequality and a Poisson limit theorem for nonoverlapping balanced subgraphs of a random graph, Random Structures Algorithms 1 (2) (1990) 231–242.
[5] M. Krivelevich, B. Sudakov, V. Vu, Covering codes with improved density, IEEE Trans. Inform. Theory 49 (7) (2003) 1812–1815.